Research Report

Approaches to the Design of Diagnostic Item Models

Edith Aurora Graf

February 2008

ETS RR-08-07

Approaches to the Design of Diagnostic Item Models

Edith Aurora Graf
ETS, Princeton, NJ

February 2008

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Quantitative item models are item structures that may be expressed in terms of mathematical variables and constraints. An item model may be developed as a computer program from which large numbers of items are automatically generated. Item models can be used to produce large numbers of items for use in traditional, large-scale assessments. But they have potential for use in other areas as well, including diagnostic assessment. In this report, I first review research on diagnostic assessment and then discuss how approaches to diagnostic assessment can inform the design of diagnostic item models.

Key words: Automatic item generation, cognitive models, diagnosis, diagnostic assessment, item modeling, item models, mathematics assessment

Acknowledgments

An earlier version of this paper was presented on April 11, 2007 at a related paper session at the 2007 annual meeting of the National Council on Measurement in Education in Chicago, Illinois. As discussants, Joanna Gorin and Derek Briggs provided valuable comments and insights. Also, I would like to thank René Lawless for chairing the session, and the authors of the related papers: William Bart, Kentaro Kato, Pam Kraus, Tara Madhyastha, Jim Minstrell, Dylan Wiliam, and Caroline Wylie. Finally, I would like to thank Isaac Bejar, Elizabeth Marquez, and Don Powers for their reviews. Any errors contained herein are the sole responsibility of the author.

The term *item model* was used by LaDuca, Staples, Templeton, and Holzman (1986) to describe classes of items that assess the same content. Bejar (2002) used the term to refer to a set of items that share a common set of structural characteristics and psychometric properties. The focus of this research is on the development of quantitative item models, the structure of which can be expressed in terms of mathematical variables and constraints. An item model may be developed as a computer program, from which many items may be automatically generated (e.g., Macready & Merwin, 1973; Singley & Bennett, 2002).

One application of item modeling is to generate large numbers of similar items for use in large-scale testing programs. Especially if items may be calibrated at the model level, this application may be an economical approach to test development. Research efforts in this direction have focused on examining the statistical comparability of model-based items (e.g., Meisner, Luecht, & Reckase, 1993; Swygert, Scrams, Thompson, & Kerman, 2006) and the evaluation of psychometric models designed to capture familial relationships among similar items (Glas & van der Linden, 2003; Sinharay & Johnson, 2005).

Ideally, item models are developed in accordance with a framework that specifies the goals for an assessment. Once they have been developed, item models should be empirically evaluated (e.g., Bejar & Yocom, 1991; Embretson & Gorin, 2001). Such an evaluation may result in revision to the item models, the underlying framework, or both. A second application of item modeling, then, is as a data collection mechanism for evaluating both the item models and the underlying framework. This paper focuses on this second application of item modeling. In particular, it considers (a) how diagnostic item frameworks may inform item model design and (b) how data collected from item models may inform underlying diagnostic frameworks.

Diagnostic Item Design

Dimensions of Diagnosis

Diagnosis can vary along a number of dimensions, including purpose, specificity, and focus. In some situations, the purpose of diagnosis is to assess mastery with respect to a set of target skills, while in others it is to identify misconceptions or procedural errors. Diagnosis can consist of high-level information, or it can highlight a particular step in a solution or protocol. The focus of diagnosis may be on an individual or on a very large group of students. The work of Brown and Burton (1978) and VanLehn (1983) focused on the identification of procedural errors, or bugs, while Minstrell (1992; 2001) used the term *facets* to refer to students' developing

ideas in a scientific domain. Some facets indicate a low level of understanding that may hinder further student learning, while others reflect an expert's view.

Depending on the purpose of the diagnosis, either a unidimensional or a multidimensional data structure may be more appropriate. If the goal is to locate a student with respect to level of understanding in a narrow domain, a unidimensional data structure may suffice. If the goal is to report performance on a constellation of skills in a broad content area, a multidimensional data structure will provide more meaningful information.

Diagnosis is often rendered by a teacher in a classroom setting, but the process can be automated by programming the rules for diagnosis into a computer-based system. In a recent report, Underwood (2007) reviewed six such systems for mathematics problem solving and identified them along several dimensions, including: How knowledge, skills, and abilities (KSAs) were represented, how levels of proficiency for the KSAs were determined, and whether common errors were identified. She found that while all six systems were designed to identify common errors, they varied with respect to how the KSAs were represented and how the KSA proficiency levels were determined.

The Unit of Diagnostic Evidence: Items Versus Item Sets

Two main approaches are typically used when the goal is to assess mastery or to identify bugs or misconceptions. In Approach 1, items are developed so that each response to each item is interpretable with respect to diagnosis. In this approach, each possible student response is linked to at least one student idea. In Approach 2, items are developed as collections so that each observation consists of a set of responses to the items in the collection. Here, the unit of analysis is a set of responses, and the response to any particular item is usually not interpreted. Bart and Williams-Morris (1990) distinguished between item diagnostic properties and test diagnostic properties. Approach 1 focuses on the former while Approach 2 focuses on the latter. An assumption made by Bart and Williams-Morris in their work was that "…tests are diagnostic to the extent to which the constituent items are diagnostic" (p. 146). In practice, diagnostic item development often combines the two approaches. Several examples follow.

Guttman and Schlesinger (1967) described how *facet design* may be used to inform the systematic construction of distractors in multiple-choice items. Their approach was based on the hypothesis that distractors that are more similar to the correct response are relatively more attractive than distractors that are less similar to the correct response. As they used the term, a

facet is a dimension along which a distractor may assume different values. Each option, including the correct response, represents a unique set of facet values. A distractor that shares a greater number of facet values with the correct response is considered more similar to it than a distractor that shares a lesser number of facet values with the correct response.

On a test of analytical ability, Guttman and Schlesinger (1967) found that distractors that were more similar to the correct response were also more attractive. They suggested that the use of systematically designed distractors potentially provides a couple of advantages: First, since each response may be differentially scored, even short tests may provide valuable information, and second, the systematic construction of distractors may make it possible to identify common student errors. In Guttman and Schlesinger's scheme, the design of each item makes use of Approach 1. The interpretation of a student's performance on a test composed of such items, with respect to facets, makes use of Approach 2.

In the context of developing diagnostic multiple-choice items for proportional reasoning, Bart, Post, Behr, and Lesh (1994) defined the properties of a *semi-dense item.* A semi-dense item has a one-to-one correspondence between each option and each cognitive rule in a *cognitive microtheory* (pp. 4–5). Bart et al. referred to the requirement that each cognitive rule should be represented among the options as *exhaustive rule set usage* (p. 4). Where diagnostic assessment development is concerned, a semi-dense item represents a useful standard, but it is an ideal that is difficult to achieve in practice. Developing semi-dense items for diagnosis is consistent with Approach 1.

The classic work of Noelting (1980), also developed in the context of proportional reasoning, primarily used Approach 2. In each task, a child was asked to compare the relative taste of two solutions, each of which contained some amount of water and some amount of orange juice. For each task the options were the same: (a) Solution A is stronger, (b) Solution B is stronger, or (c) the two solutions have the same strength. Noelting designed his tasks to diagnose childrens' levels of development, and the tasks were constructed to be interpreted as a set. The tasks were ordered according to the level of reasoning required. Easier tasks (for example, those with solutions with different amounts of water but the same amounts of orange juice) required only qualitative reasoning, while the hardest tasks required students to compare ratios by finding a common denominator.

Noelting (1980) used a scalogram analysis (see Guttman, 1944) to order the items by difficulty. To define the developmental stages, he grouped children who were able to answer items at similar levels of difficulty and labeled each stage in accordance with a Piagetian framework. A child's developmental stage was interpreted with respect to the cognitive demands of the most difficult items he or she was able to answer. Noelting verified six of the seven stages he identified through the use of confirmatory factor analysis.

Bart and Williams-Morris (1990) analyzed Noelting's items with respect to two of the properties they had defined for a semi-dense item: *response interpretability* and *response discrimination*. Response interpretability refers to the requirement that each response be interpretable by a cognitive rule; response discrimination refers to the requirement that each response be interpreted by one and only one cognitive rule. They developed item indices to measure each of these properties. The proportion of responses that could be interpreted by at least one cognitive rule indicated the response interpretability of an item. Response discrimination for an item was calculated as the average discrimination across responses, where the discrimination for each response was inversely proportional to the number of cognitive rules that would result in that response. In the case where the response was not interpretable, discrimination was zero.

An application of Approach 2 is concerned with detecting patterns of responses that reflect strengths and weaknesses for a set of target skills. Tatsuoka, Corter, and Tatsuoka (2004) and Birenbaum, Tatsuoka, and Yamada (2004) used the rule space method to compare student mastery patterns in mathematics, across countries. Students' responses to eighth-grade mathematics items from the Third International Mathematics and Science Study-Repeat (TIMSS-R) were used to infer patterns of mastery with respect to a set of attributes, and the mean attributes from different countries were compared. It was found that U.S. students were relatively weaker with respect to geometry content and logical reasoning process attributes. In this example, diagnosis focused on attributes which were interpreted with respect to a set of items, and the focus was on the performance of large subgroups.

In a recent chapter, Luecht (2007) used a data augmentation approach to explore whether distractor selections from multiple-choice items could be used to yield more meaningful diagnostic subscores. The correlations among the augmented subscores were extremely high, however, so there was no evidence that information on multiple skills had actually been

extracted. The high correlations may have been due in large part to the inherent unidimensionality of the tests under study. Nevertheless, Luecht concluded that the approach may still be informative in situations where collateral multidimensional information can be extracted. This suggests that for tests that assess multiple skills by design, distractors may still provide additional useful diagnostic information.

The DIAGNOSER (Hunt & Minstrell, 1994; Levidow, Hunt, & McKee, 1991) is a computer program that was developed to identify student facets. In the DIAGNOSER, each response option is linked to a facet, or student idea. The use of the term facet is somewhat different in the works of Minstrell (1992; 2001) and in the work of Guttman and Schlesinger (1967). Guttman and Schlesinger used the term to refer to a dimension of an item (in particular, a dimension of a distractor). As Minstrell used the term, a facet is a student idea linked to a particular level of understanding, and the idea is almost always content specific. Minstrell originally developed facets in the context of high school physics instruction, but they have since been extended to other domains, like statistics (Schaffner et al., 1996). To a large extent, DIAGNOSER items represent an application of Approach 1—each response option is linked to a facet, which represents a student idea. Each response constitutes an observation that provides evidence with respect to specific difficulties in understanding, as well as level of mastery (since facets are ordered from more expert to more problematic).

DIAGNOSER items make use of Approach 2 as well, however. *Facet clusters* (Minstrell, 2001, p. 420) consist of conceptually related facets that are applicable to a unit of instruction. Since the number of facets in a cluster typically exceeds the number of ideas that is relevant to a particular question, exhaustive rule set usage is not a requirement for the development of a particular item. A student's response to a particular question thus provides evidence regarding a subset of facets (Approach 1), while responses to several questions provide evidence regarding standing with respect to a cluster (Approach 2).

Briggs, Alonzo, Schwab, and Wilson (2006) developed a set of diagnostic multiple-choice items to assess levels of understanding in *Earth in the Solar System*. Earth in the Solar System content is concerned with how objects in our solar system move to produce cyclical changes, for example between day and night or across the seasons. They first developed construct maps in the domain, which provided descriptions of evidence for different levels of understanding and associated them with score levels. The construct maps were based on prior

research findings of student understanding in the domain and were also informed by national standards for Grades 5 and 8. The items were designed so that each option corresponded to a developmental level from the construct map. Thus Briggs et al. incorporated theories of learning development and misconceptions, as well as requirements from the national standards into the design of their items. Because each option corresponds to a developmental level, Briggs et al. make strong use of Approach 1. They suggested that Wilson's ordered partition model, a psychometric model that allows for the interpretation of particular responses, could be used to support a large-scale implementation of an assessment based on the items they developed.

The use of Approach 2 is also implied by the Briggs et al. (2006) work. Their construct map provided a single developmental scale that pertained to concepts covered in Grades 5 through 8. They developed items at the Grade 5 level and items at the Grade 8 level. Items developed for the Grade 5 level did not include options that represented the highest level from the construct map, which corresponded to the level of understanding expected from students at eighth grade (as reflected in the national standards documents). Items developed for the Grade 8 level could include options that corresponded to developmental levels usually observed at the fifth grade, however. This scheme potentially allows for vertical comparisons across grades in at least one direction—for example, it is possible to identify an eighth grader operating at a level of development normally associated with the fifth grade.

Smith, Wiser, Anderson, and Krajcik (2006) also provided an example of how an assessment can be developed in accordance with models of student development, or *learning progressions*. According to Smith et al., learning progressions are "…descriptions of successively more sophisticated ways of reasoning within a content domain based on research syntheses and conceptual analyses…" (p. 1), that "…should be organized around central concepts and principles of a discipline (i.e., its *big ideas*)…" (p. 2). In their work, Smith et al. proposed a learning progression for student understanding of matter and atomic molecular theory. For each of three grade bands (K-2, 3-5, and 6-8), levels of understanding were framed with respect to a common set of big ideas. In general, the levels of understanding are more nuanced and complete in the higher grade bands. For example, one component of a big idea is that students should understand that "mass and weight are conserved across a broad range of transformations" (Smith et al., Figure 1, p. 15). At the K-2 level, students are expected to understand that physical transformations such as breaking into pieces will conserve weight and

the "amount of stuff." By Grades 6-8 however, they are expected to understand that "mass and weight are conserved in physical and chemical changes because atoms are neither created nor destroyed" (Smith et al., Figure 1, p. 15).

Smith et al. (2006) also discussed how *learning performances* can provide evidence for student understanding of both scientific content and practice. Assessment items can be developed that elicit such evidence, and they developed several sample items for each grade band. Several of the tasks are diagnostic in the sense that particular learning performances are indicative of how far an idea has developed with respect to a learning progression. For example, in a sample task Smith et al. developed for Grades 3-5, students were first shown two solid rectangular prisms that were equal in volume and weight. They were asked to consider whether the two prisms could have been made of the same material. Then, they were shown two solid prisms that had equal volumes, but one prism was heavier than the other. Again, they were asked to consider whether the two prisms could have been made of the same material. Smith et al. suggested that students who have an early concept of density would answer yes to the first question but no to the second, whereas students who have not yet developed this concept might answer yes to both questions.

Extensions to Diagnosis

The discussion in the preceding section suggests that methods of diagnostic assessment have been extended to support both developmental and standards-based interpretations. Minstrell (2001, p. 148) proposed that a system that provides descriptions of students' understanding should meet the following requirements: It should be grounded in research, it should specify learning expectations, it should characterize a progression from naïve ideas to the learning expectations, it should identify student difficulties, and it should be tractable to both theoreticians and practitioners in the field. The learning expectations might be state or national standards, or they might be competencies established as goals for a particular assessment. The goals for diagnostic assessment have expanded with the advent of standards-based reform and the recognition that understanding the trajectory of development in the target domain is necessary to help students reach learning expectations.

Early work in diagnostic assessment focused on the identification of bugs and misconceptions. Errors have been identified that occur with very high frequency. For example, in proportional reasoning, the *incorrect addition strategy* (Hart, 1984) is very common. Research

has also examined the stability of bugs and misconceptions across contexts and within individuals. The general finding has been that students are more or less prone to certain errors given changes in context, and that individuals are not necessarily consistent in their reasoning. For example, Payne and Squibb (1990) found that the same student will tend to use different *malrules*, even on very similar algebra items. Madhyastha, Hunt, Kraus, and Minstrell (2006) found that students did not generally endorse physics facets consistently, although consistency improved with both instruction and math ability. This result appears to agree with an observation made by Payne and Squibb that although students tend to be inconsistent in their application of mal rules, it is typically easier to diagnose students with greater levels of algebra skill.

These findings suggest a course of action: Identification of common errors and misconceptions may be extremely useful for the purpose of guiding classroom instruction. Where making recommendations for an individual is concerned, consistency of responses is also important to evaluate. Common misconceptions may be addressed effectively via a class discussion. A student who responds inconsistently may require individual attention from the teacher. To the extent that inconsistent responses reflect weak understanding, direct, basic instruction may be necessary.

Not all diagnostic assessment may be considered formative. Formative assessment can positively impact student learning, however (e.g., Black & Wiliam, 1998; Wiliam, Lee, Harrison, & Black, 2004). An example of how diagnostic items may be used formatively in the classroom is described by Ciofalo and Wylie (2006) and Wylie and Wiliam (2006)—teachers used diagnostic items individually to guide the course of instruction in real time.

As the intended applications for diagnostic assessment expand, it will be necessary to extend and combine approaches to diagnosis in novel ways. This requirement was forecasted succinctly by Bejar (1984, p. 175), "…the traditional approach to the specification of content in terms of static taxonomies may not be appropriate given the dynamic and sequential nature of diagnostic assessment."

Item Type Considerations

Not surprisingly, different item types are more or less suited for different aspects of diagnosis. Multiple-choice items are inexpensive and easy to score, and scoring accuracy is very high. Students can respond to them quickly, and are likely to interpret them correctly because the options cue the expected nature of the response. Multiple-choice items are ideal in situations

where the goal is to determine whether a student can recognize the correct response and reject incorrect but viable alternatives. Especially when students are at the early stages of learning, multiple-choice items may be valuable for detecting the presence of partial knowledge.

The research summarized earlier supports the idea that diagnostic multiple-choice items may be used to efficiently identify student misconceptions and their associations with stages of development. Multiple-choice items have the disadvantage that a response (whether or not it is correct) may reflect a guess rather than a firmly held belief. Mathematics items in particular may be susceptible to *backsolving* (for further discussion of this strategy, see Braswell & Jackson, 1995; Bridgeman, 1993). Multiple-choice items do not provide specific evidence regarding a student's solution procedure or the details of his or her reasoning. This is not so much a weakness as a limitation. Multiple-choice items are highly efficient for collecting diagnostic information, particularly when the options have been constructed in accordance with a theory of learning in the domain.

Both Bridgeman (1993) and Katz, Bennett, and Berger (2000) suggested that it may be useful to develop multiple-choice items that include common incorrect responses as distractors—and the recent work of Briggs et al. (2006) suggests that ordered multiple-choice items may be as reliable as traditional multiple-choice items while providing greater diagnostic evidence. It is also worth pointing out that not all items that may be scored as multiple-choice must be presented in the standard format, with a stem followed by several options.

Constructed-response items serve a complementary role in diagnostic assessment. They are more representative of real-world tasks than multiple-choice items and may discourage guessing. Constructed-response items of the "show all work" type provide highly specific information about the details of students' solution methods. The scoring accuracy for constructed-response items is generally lower than for multiple-choice items, but some in mathematics can be scored quite accurately, even compared to multiple-choice. For example, Bennett, Steffen, Singley, Morley, and Jacquemin (1997) found very high accuracy rates for the mathematical expressions (ME) response type when users entered expressions on the computer. In their study, the accuracy rate for scoring the ME type was 99.62% with real response data. Scoring accuracy for responses that were designed to be difficult to score was lower (70%), but these responses were complex. By comparison, scoring accuracy for multiple-choice items is about 99.95% (J. McDonald, as cited in Bennett et al.).

Constructed-response items have the disadvantage that they can be expensive to score (even when scored automatically), and students usually take more time to respond to them. This means they can consume valuable time on an assessment or in the classroom. Nevertheless, because they are less amenable to backsolving, constructed-response items place different cognitive demands on the student—he or she must generate a response without the support of the options. Constructed-response items can also provide more evidence regarding the details of solution. Finally, they may be somewhat faster to develop than multiple-choice items with conceptually based distractors. This is not to suggest that a constructed-response item is easy to develop—additional care must be taken in the wording of the stem since there are no options to support its interpretation.

In sum, multiple-choice and constructed-response items provide different but complementary evidence for diagnosis. In developing a diagnostic assessment, it may be very useful to use these item types together.

Supporting Frameworks

The outcome aspect of validity refers both to test interpretation and test use (Messick, 1989, p. 20). Both are important to consider in the design and implementation of diagnostic assessment. Ideally, the construct is fully specified before item development begins. Increasingly, assessments are developed in accordance with *cognitive models* (e.g., Gorin, 2006). Gorin distinguished between more general cognitive models developed for the purpose of construct definition and more specific cognitive models specific to an item or item type. Although this paper focuses primarily on applications of the latter, it is assumed that cognitive models at both levels should be developed, and that a cognitive model that explains the construct should be drafted first. Following each round of empirical validation or evaluation for logical consistency between cognitive models at different levels, the models are typically refined.

Diagnostic assessment can vary along all of the dimensions previously discussed, but there are a number of frameworks that are sufficiently general to guide the design of the underlying cognitive models. Examples of such frameworks include a framework for developing a cognitively diagnostic assessment (CDA; Nichols, 1994), evidence-centered design (ECD; e.g., Mislevy, Steinberg, & Almond, 2003), and the cognitive design system approach (Embretson & Gorin, 2001; Embretson, 1999). Nichols' framework specified five steps critical to the development of a CDA, as follows:

1. substantive theory construction
2. design selection
3. test administration
4. response scoring
5. design revision

Step 1 refers to the construction of a general cognitive model for the purpose of construct definition, as described in Gorin (2006). This step is essential, though the discussion here is primarily concerned with Step 2, design selection. Nichols provided the following description of the design selection step:

> In this step, the test developer selects the observation and measurement designs. The selection is informed by the substantive base constructed in Step 1. Subsequently, the test developer constructs items or tasks that will be responded to in predictable ways by test takers with specific knowledge, skills, and other characteristics identified as important in the theory. The procedure for constructing assessments is the operationalization of the assessment design. (p. 587)

In ECD, the student model is used to define the construct, while the task models specify features of tasks. Cromley and Mislevy (2004) extended ECD by incorporating misconceptions into a template structure. Shute, Graf, and Hansen (2005) described how ECD was applied to the development of the Adaptive Content with the Evidence-Based Diagnosis (ACED) system, a computer-based diagnostic assessment that focused on mathematical sequences suitable to assess at the eighth-grade level.

In the introductory chapter to the edited book, *Cognitive Diagnostic Assessment for Education*, Leighton and Gierl (2007) reviewed research that has led to the development of CDA, including Nichols' (1994) framework. The goals of CDA are to provide evidence about students' thought processes in a domain and to characterize their strengths and weaknesses. Leighton and Gierl concluded that developing a CDA requires both a strong cognitive theory and a means for scientifically validating the theory.

Item Models That Accommodate Different Diagnostic Dimensions and Item Types

The usual method for designing item models for use in large-scale summative assessments is to select a set of previously calibrated *source items.* Each source item is parameterized; components of the item (both numbers and strings) are replaced with variables. The details for how to generalize a quantitative source item to an item model are described in Graf, Peterson, Steffen, and Lawless (2005). The parameterized source item then serves as a template for automatically generating *instances* (Bejar, 2002), or discrete items, from the item model. Instances are generated by instantiating variables in the template with particular values. This method of automatic item generation makes use of what is referred to as *replacement-set procedures* (e.g., Millman & Westman, 1989) and is now almost standard. Instances generated from an item model that relies solely on replacement-set procedures often appear quite similar. There has been research to explore advances in automatic item generation to allow for more abstract forms of model description (Deane, Graf, Higgins, Futagi, & Lawless, 2006; Deane & Sheehan, 2003; Higgins, Futagi, & Deane, 2005).

Regardless of the method used for generation, item model development should not be item centric—rather, item models should be designed in accordance with more general schemas (for examples of this approach, see Enright, Morley, & Sheehan, 2002 and Singley & Bennett, 2002). Although software that uses replacement-set procedures for automatic generation cannot be used to characterize item models at abstract levels of description, it is still possible to design item *families*, where each family is an organized collection of item models that represents variations on a common schema.

Developing Diagnostic Item Models

In this part of the paper, I consider how the research summarized earlier can inform principles for developing diagnostic item models. The discussion of item types suggests that for diagnostic assessment, there may be a need for either multiple-choice or constructed-response item models, depending on the setting and the kinds of evidence that need to be collected. Variables may be included in any part of an item model, including the stem, the key, and the options, if there are any. Constructed-response item models consist of a *stem model* and a *key model*, while multiple-choice item models consist of a stem model, a key model, and *distractor models*.

Three example item models are shown in Table 1; one item model is shown in each row. Since multiple-choice items and constructed-response items serve potentially complementary roles in diagnostic assessment, these examples have been deliberately structured so that they could generate either multiple-choice or constructed-response instances. The first column of Table 1 provides a high level description of the problem setting and goal. The second column shows the parameterized presentation frame, or template, for the item model. The third column defines the variables and constraints used in the model, and the fourth column shows an instance that could be generated from the model.

The final column in Table 1 shows links to the National Council of Teachers of Mathematics (NCTM) curriculum focal points (CFPs; 2006), which were designed to highlight essential content themes in preK-8 mathematics. The themes addressed by the examples are given in the last column. The first item model example is a very basic skills question that asks the student to find the area of a circle given its radius. When an instance is generated, the variable that represents the radius is replaced with a numeric value. The instances generated from this model are suitable for seventh grade. Some common errors that might be expected are confusing the formula for circumference with the formula for area (Error 1) or a confusion between radius and diameter (Error 2). These *error models* might be represented as distractors in a multiple-choice instance, or they might be used to interpret incorrect responses to a constructed-response instance. Morley, Lawless, and Bridgeman (2005) took a similar approach in their modeling of answer-choice rationales.

The second item model example in Table 1 is a word problem, and it requires that the student represent a situation as a mathematical expression. Of the components that have been parameterized, only two are likely to influence item difficulty: The relation between the original distance between the arms of the compass and the final distance, and the value for this change. As with the first item model example, the Error 1 and Error 2 models refer to confusions between area and circumference, and radius and diameter, respectively. Error 3 is more conceptual: A student who makes this error probably has an insufficient model of the problem and has likely neglected to consider the original width between the arms of the compass in finding the expression. The instances generated from this model might also be suitable for seventh grade.

Table 1

A Family of Three Diagnostic Item Model Examples

Description of model	Presentation frame	Variables and constraints	Instance	Link to NCTM CFPs
Given the radius of a circle, find its area.	What is the area of a circle with radius ***r***?	***r*** is an integer between 3 and 15, inclusive. $Key: \pi \boldsymbol{r}^2$ $Error1: 2\pi \boldsymbol{r}$ $Error2: \pi(2\boldsymbol{r})^2$	What is the area of a circle with radius 4? $Key: 16\pi$ $Error1: 8\pi$ $Error2: 64\pi$	In Grade 7, students use concepts from geometry, measurement, and algebra to model and solve a large variety of problems, including those involving finding areas of circles (National Council of Teachers of Mathematics [NCTM], 2006, p. 19)
Use an expression to represent what happens to the area of a circle if a constant is added to or subtracted from its radius.	***Name***'s compass is set so that the distance between the ends of its arms is length *r* units. ***Name*** draws a circle. ***Name.PN*** adjusts one of the arms so that the distance between the ends is ***s*** units ***relation*** than it was before, and draws a second circle. Write an expression that represents the area of the second circle.  	***Name*** is a string representing a name. ***Name.PN*** is a string representing a pronoun that agrees with ***Name.*** ***s*** is a integer between 1 and 15, inclusive. ***relation*** is a string that is either "greater" or "less." ***relation.sign*** is a string that varies with ***relation***. If ***relation*** is "greater," then ***relation.sign*** is "+"; otherwise ***relation.sign*** is "–". $Key: \pi(r\ relation.sign\ \boldsymbol{s})^2$ $Error1: 2\pi(r\ relation.sign\ \boldsymbol{s})$ $Error2: \pi(2(r\ relation.sign\ \boldsymbol{s}))^2$ $Error3: \pi \boldsymbol{s}^2$	Anita's compass is set so that the distance between the ends of its arms is length *r* units. Anita draws a circle. She adjusts one of the arms so that the distance between the ends is 3 units greater than it was before, and draws a second circle. Write an expression that represents the area of the second circle. $Key: \pi(r+3)^2$ $Error1: 2\pi(r+3)$ $Error2: \pi(2(r+3))^2$ $Error3: 9\pi$	

(Table continues)

Table 1 (continued)

Description of model	Presentation frame	Variables and constraints	Instance	Link to NCTM CFPs
Given two points in the Cartesian plane, one on the circumference of a circle and another at its center, find the area of the circle.	The coordinates of two points in the *xy*-plane are ($\boldsymbol{X_a}$, $\boldsymbol{Y_a}$) and ($\boldsymbol{X_b}$, $\boldsymbol{Y_b}$). Find the area of a circle that has one of these points as its center and the other on its circumference.	$\boldsymbol{X_a}$ is a nonzero integer between -5 and 5, inclusive. $\boldsymbol{Y_a}$ is a nonzero integer between -5 and 5, inclusive. $\boldsymbol{\Delta X}$ is a nonzero integer between -9 and 9. $\boldsymbol{\Delta Y}$ is a nonzero integer between -9 and 9. $\boldsymbol{X_b} = \boldsymbol{X_a} + \boldsymbol{\Delta X}$ $\boldsymbol{Y_b} = \boldsymbol{Y_a} + \boldsymbol{\Delta Y}$ $\boldsymbol{X_b} \neq 0; \boldsymbol{Y_b} \neq 0$ $\boldsymbol{\Delta X}^2 + \boldsymbol{\Delta Y}^2 > 4$ $Key: \pi\left(\boldsymbol{\Delta X}^2 + \boldsymbol{\Delta Y}^2\right)$ $Error1: \pi\left(\boldsymbol{X_a}^2 + \boldsymbol{Y_a}^2\right)$ or $\pi\left(\boldsymbol{X_b}^2 + \boldsymbol{Y_b}^2\right)$ $Error2: 2\pi\sqrt{\boldsymbol{\Delta X}^2 + \boldsymbol{\Delta Y}^2}$ *Error* 3: Can't solve the problem because I don't know which point is at the center and which is on the circumference.	The coordinates of two points in the *xy*-plane are (-3, 1) and (-1, -6). Find the area of a circle that has one these points as its center and the other on its circumference. $Key: 53\pi$ $Error1: 10\pi$ or 37π $Error2: 2\sqrt{53}\pi$ *Error* 3: Can't solve the problem because I don't know which point is at the center and which is on the circumference.	"…They apply the Pythagorean theorem to find distances between points in the Cartesian coordinate plane to measure lengths…" (NCTM, 2006, p. 20)

Note. Item model variables are shown in bold. NCTM = National Council of Teachers of Mathematics; CFP = curriculum focal points.

The final item model example in Table 1 is designed for eighth grade, and builds on the knowledge required by the first item model. Now, instead of finding the area of a circle for which the radius is given, students must apply the Pythagorean theorem to find the radius of a circle in the Cartesian plane. Error 1 represents the condition where a student misinterprets the radius as the distance between the origin and one of the two points. Error 2 represents confusion between area and circumference. Error 3 is conceptual and suggests a recommendation to the student: It is helpful to draw a diagram and to consider different possible cases.

It is interesting to consider the different kinds of evidence that the last example might provide when presented in either multiple-choice or constructed-response format. The student may or may not recognize that in order to find the area, it is not necessary to calculate the radius by taking the square root of $\left(\Delta X^2 + \Delta Y^2\right)$. If the item is presented in constructed-response format and the student shows work, it will be evident from the response whether the student has calculated the radius by taking the square root and then squared it again to find the area. When the instances are presented in multiple-choice format, a student's solution method will not be directly observable. It is predicted, however, that students who calculate the radius will find instances for which the radius squared is not a perfect square considerably more difficult. Students who do not calculate the radius should find both categories of instances relatively comparable in difficulty.

The final example may also be difficult due to the vocabulary. The term *circumference* can refer either to a measure or a boundary; in this case the second definition is intended. If the meaning of the word circumference is not clear from the context, the student may have difficulty with the task. Like other English words, mathematics vocabulary words often have multiple meanings, and this can be challenging for students (Thompson & Rubenstein, 2000). The final example could be reworded as follows: Given two points in the Cartesian plane, one on the circle and another at its center, find the area of the circle. This wording potentially poses different difficulties, however: Although the wording *on the circle* is technically correct, many students may confuse *circle* with *circular region*. A student who mistakenly interprets *circle* to mean *circular region* may not know how to answer the question, even if he or she understands the concepts involved. A benefit to a model-based approach to test development is that one can generate instances with different wording to evaluate the impact of linguistic features, such as

vocabulary, on difficulty. Approaches to automatic item generation that go beyond replacement-set procedures and allow for alternate phrasings may be especially useful for this purpose.

At this point, quite a bit of research has systematically explored the impact of various item model variables on difficulty (e.g., Bejar, 1993; Bejar & Yocom, 1991; Embretson, 1999; Enright et al., 2002; Graf et al., 2005; Newstead, Bradon, Handley, Evans, & Dennis, 2002). The item models shown in Table 1 form a natural family. The first model is a basic skills problem; the second model requires representation. The last model builds on the knowledge required by the first model. Because the errors have also been modeled in these examples, misconceptions and/or bugs are consistently represented across instances. From an experimental design standpoint, this approach is valuable, because it allows for a strong test of misconceptions, disambiguated from a particular context. The more general the item model, the stronger the test.

Analyzing and Validating Diagnostic Item Models

The previous discussion describes how diagnostic items can be generalized and how errors can be modeled. As with discrete item development, one could apply the semi-dense framework of Bart et al. (1994) to guide the development of diagnostic item models. Table 1 suggests how this might occur. Modeling errors in accordance with cognitive rules ensures that at least some possible responses to all instances will be interpretable. So modeling errors may enhance the response interpretability of all generated instances. The response discrimination property may also inform the development of item models. The item model author must consider the discrimination of each response to each instance. As values change across instances, it is possible that while some instances include responses that are interpretable by one and only one cognitive rule, other instances include responses that are interpretable by more than one cognitive rule. Consider the first item model in Table 1—if the radius were allowed to assume a value of two units, the response "4π" might be explained by either the key or Error 1. Thus the response discrimination for this instance would be lower than for other instances. Typically, constraints may be introduced to ensure that all instances have comparable levels of response discrimination. If responses are modeled in accordance with cognitive rules, the response discrimination property can be satisfied for all instances by setting constraints so that no two responses are equivalent.

The examples in Table 1 also show how meaningful sets of item models can be constructed. A student's responses to instances generated from all three models in Table 1 provide more information about his or her understanding concerning the area of a circle than his or her responses to instances generated from a single item model from Table 1. Students who can successfully respond to instances generated from the first item model but not to instances generated from the second or third item models have learned how to calculate area given the radius but cannot extend this knowledge to new situations or applications. Most students who can successfully respond to an instance generated from the second or third model will probably also respond correctly to an instance generated from the first model; but it might be expected that different students would have more or less difficulty with instances generated from either the second or the third models, depending on whether they are relatively more skilled at representing situations as algebraic expressions or calculating distances in the Cartesian plane. Students who can successfully respond to instances generated from all three models have demonstrated a relatively broader scope of understanding where the relationship between radius and area is concerned.

The examples in Table 1 are just for the purpose of illustration. Ideally, errors are modeled in accordance with findings from the literature and validated through empirical study. The items developed as part of the Diagnostic Items for Math and Science (DIMS) project (Ciofalo & Wylie, 2006; Wylie & Wiliam, 2006) were based on a review of the literature and the expertise of practitioners. A recent pilot study investigated the comparability of instances generated from item models based on four different DIMS questions for eighth-grade mathematics (Graf, Ohls, Klag, & Wylie, 2006). Ten instances were generated from each model, and students responded to each instance. One finding was that while the instances generated from these item models were not quite isomorphic with respect to difficulty, the misconceptions endorsed across instances were reasonably consistent. This finding suggests that some distractors were attractive due to the misconceptions they represented, rather than to the particular context in which they appeared. It has been suggested that an item modeling approach to automatic item generation may enhance validity (Bejar, 1993; Bejar & Yocom, 1991). It is likely that it may be used as a research methodology for exploring validity in diagnostic assessment contexts as well.

Summary

The first part of this report discussed dimensions of diagnosis and approaches for creating and analyzing diagnostic items. Diagnostic items can be developed so that they may be interpreted individually or as part of a set; these approaches are often used together. Ideally, the development of a diagnostic assessment is informed by both a higher level cognitive model that defines the construct and a more specific cognitive model that distinguishes the roles of specific item types in how the construct is assessed. Multiple-choice and constructed-response formats serve complementary roles in diagnosis, and each is best suited for different purposes.

The second part of the paper focused on the implications of this discussion for the development and validation of diagnostic item models. It was described how item models can be designed to accommodate both approaches to diagnostic item development, and that they may be used to generate instances in different response formats. Finally, it was suggested that item modeling is a potentially valuable research tool in the validation of diagnostic assessment.

References

Bart, W. M., Post, T., Behr, M. J., & Lesh, R. (1994). A diagnostic analysis of a proportional reasoning test item: An introduction to the properties of a semi-dense item. *Focus on Learning Problems in Mathematics*, *16*(3), 1–11.

Bart, W. M., & Williams-Morris, R. (1990). A refined item digraph analysis of a proportional reasoning test. *Applied Measurement in Education, 3*(2), 143.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement, 21*(2), 175–189.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen (Ed.), *Test theory for a new generation of tests* (pp. 323–357). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–218). Mahwah, NJ: Lawrence Erlbaum Associates.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15*(2), 129–137.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*(2), 162–176.

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Between-countries and within-country comparisons of eighth graders' mathematics performance. *Studies in Educational Evaluation, 30*, 151–173.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, Principles, Policy & Practice, 5*, 7–74.

Braswell, J. S., & Jackson, C. A. (1995, April). *An introduction of a new free-response item type in mathematics.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bridgeman, B. (1993). *A comparison of open-ended and multiple-choice question formats for the quantitative section of the Graduate Record Examinations General Test* (ETS Research Rep. No. RR-91-35). Princeton, NJ: ETS.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33–63.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*(2), 155–192.

Ciofalo, J. F., & Wylie, E. C. (2006, January 10). *Using diagnostic classroom assessment: One question at a time.* Retrieved December 3, 2007, from the Teachers College Record Web site: http://www.tcrecord.org/content.asp?contentid=12285

Cromley, J. G., & Mislevy, R. J. (2004). *Task templates based on misconception research* (CSE Rep. No. 646). College Park, MD: University of Maryland.

Deane, P., Graf, E. A., Higgins, D., Futagi, Y., & Lawless, R. R. (2006). *Model analysis and model creation: capturing the task-model structure of quantitative item domains* (ETS Research Rep. No. RR-06-11). Princeton, NJ: ETS.

Deane, P., & Sheehan, K. M. (2003). *Automatic item generation via frame semantics: Natural language generation of math word problems*. Unpublished manuscript, ETS, Princeton, NJ.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407–433.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49–74.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*(4), 247–261.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21–35.

Graf, E. A., Ohls, S., Klag, D., & Wylie, E. C. (2006, August 24). *Developing and analyzing item models based on diagnostic items for math and science (DIMS).* Presentation at the ETS Assessment Innovations Seminar Series, Princeton, NJ.

Graf, E. A., Peterson, S., Steffen, M., & Lawless, R. (2005). *Psychometric and cognitive analysis as a basis for the design and revision of quantitative item mode* (ETS Research Rep. No. RR-05-25). Princeton, NJ: ETS.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139–150.

Guttman, L., & Schlesinger, I. M. (1967). Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement, 27*(3), 569–580.

Hart, K. (1984). *Ratio: Children's strategies and errors.* Windsor, England: NFER-Nelson.

Higgins, D., Futagi, Y., & Deane, P. (2005). *Multilingual generalization of the ModelCreator software for math item generation* (ETS Research Rep. No. RR-05-02). Princeton, NJ: ETS.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51–74). Cambridge, MA: The MIT Press.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*(1), 39–57.

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modelling procedure for constructing content-equivalent multiple choice questions. *Medical Education, 20*, 53–56.

Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 3–18). New York: Cambridge University Press.

Levidow, B. B., Hunt, E. B., & McKee, C. (1991). The DIAGNOSER: A HyperCard tool for building theoretically based tutorials. *Behavior Research Methods, Instruments, and Computers, 23*(2), 249–252.

Luecht, R. M. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 319–339). New York: Cambridge University Press.

Macready, G. B., & Merwin, J. C. (1973). Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement, 33*, 351–360.

Madhyastha, T., Hunt, E., Kraus, P., & Minstrell, J. (2006, April). *The relationship of coherence of thought and conceptual change to ability.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Meisner, R., Luecht, R., & Reckase, M. D. (1993). *The comparability of the statistical characteristics of test items generated by computer algorithms* (ACT Research Rep. No. 93-9). Iowa City, IA: ACT.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.

Millman, J., & Westman, R. S. (1989). Computer-assisted writing of achievement test items: Toward a future technology. *Journal of Education Measurement, 26*(2), 177–190.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110–128). Kiel, Germany: IPN.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley (Ed.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 415–443). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Morley, M. E., Lawless, R. R., & Bridgeman, B. (2005). Transfer between variants of mathematics test questions. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 313–336). Greenwich, CT: Information Age Publishing.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics.* Retrieved September 12, 2006, from http://www.nctm.org/focalpoints/downloads.asp

Newstead, S., Bradon, P., Handley, S., Evans, J., & Dennis, I. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In S. H. Irvine & P.

C. Kyllonen (Eds.), *Item generation for test development* (pp. 35–51). Mahwah, NJ,: Lawrence Erlbaum Associates.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*(4), 575–603.

Noelting, G. (1980). The development of proportional reasoning and the ratio concept: Part I - Differentiation of stages. *Educational Studies in Mathematics, 11*(2), 217–253.

Payne, S. J., & Squibb, H. R. (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science, 14*(3), 641–642.

Schaffner, A., Madigan, D., Hunt, E., Graf, E. A., Minstrell, J., & Nason, M. (1996, July). *Benchmark lessons and the World Wide Web: Tools for teaching statistics.* Paper presented at the international conference on learning sciences, Evanston, IL.

Shute, V. J., Graf, E. A., & Hansen, E. G. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. M. Pytlikzillig, R. H. Bruning, & M. Bodvarsson (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 169–202). Greenwich, CT: Information Age Publishing.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine (Ed.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Lawrence Erlbaum Associates.

Sinharay, S., & Johnson, M. (2005). *Analysis of data from an admissions test with item models* (ETS Research Rep. No. RR-05-06). Princeton, NJ: ETS.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, *4*(1&2), 1–98.

Swygert, K. A., Scrams, D. J., Thompson, L. E., & Kerman, D. E. (2006). *An evaluation of the impact of cloning on item parameters* (Computerized Testing Rep. No. 99-08). Newtown, PA: Law School Admission Council.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901–926.

Thompson, D. R., & Rubenstein, R. N. (2000). Learning mathematics vocabulary: Potential pitfalls and instructional strategies. *Mathematics Teacher, 93*(7), 568–574.

Underwood, J. (2007). *A critical evaluation of automated diagnosis and instruction for mathematics problem-solving* (ETS Research Rep. No. RR-07-28). Princeton, NJ: ETS.

VanLehn, K. (1983, August). *Human procedural skill acquisition: Theory, model and psychological validation.* Paper presented at the American Association for Artificial Intelligence, Los Altos, CA.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment of learning: Impact on student achievement. *Assessment in Education, 11*(1), 49–65.

Wylie, E. C., & Wiliam, D. (2006, April). *Diagnostic questions: Is there value in just one?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.